Topics in Statistical Computing with R

Ian McLeod

Western University

University of Waterloo, Systems Design Seminar

Talk website

My talk is based on the article, "Time Series Analysis with R" This paper as well as the R code for all tables and scripts is available from my website:
`http://www.stats.uwo.ca/faculty/aim/tsar/`
Reference: McLeod, A. I., H. Yu & E. Mahdi (2012). Time Series Analysis with R. In *Time Series Analysis: Methods and Applications* Chapter 23 (pp. 661-712) in Handbook in Statistics, Volume 30, Edited by T. S. Rao, S. S. Rao and C. R. Rao. ISBN: 978-0-444-53858-1. Elsevier.
Entire book *Time Series Analysis: Methods and Applications* may be downloaded from the above website.

Quantitative Programming Environment

High level and *high quality programming similar to MatLab featuring:*

- *scripts, functions, packages*
- *interface to C, Fortran, etc.*
- *use multicore PC*

Advantages over MatLab:

- *it is free!*
- *open source, CRAN*
- *runs on SharcNet*
- *ideal for teaching and research*
- *python and R are widely used by Google*

More features

R is more specialized for data science:

- graphical methods and data visualization
- high quality statistical functions
- natural object-oriented environment
- formula language for specifying models
- reproducible research and Sweave
- Excel, SPSS and *Mathematica* incorporate R

R Books

Many books in statistics that are published are accompanied by R packages published on CRAN. Such packages may contain interesting datasets, vignettes and examples as well as R functions implements methods.

- *Introduction to R* from CRAN and other documents
- many statistics textbooks, elementary to advanced
- *Use R!* from Springer
- *The R Series* from Chapman Hall/CRC

Refereed Journals

The authors of many methodological research papers published in refereed statistics journals often put Packages on CRAN in the hope of increasing the impact of their research.

- *Journal of Statistical Software*
- The R Journal
- *Journal of Environmental Statistics*

Web sites

- CRAN
- CRAN views
- CRAN R FAQ
- R-help
- Rmetrics
- Bioconductor
- Revolution analytics

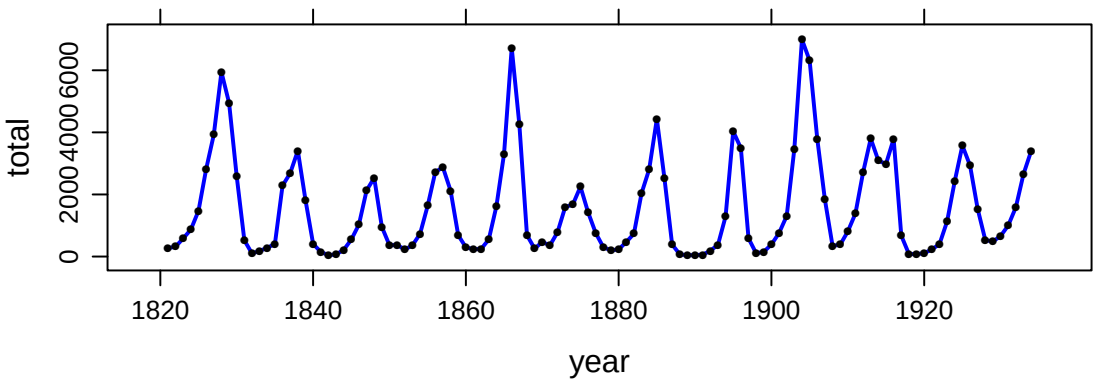
R Editors I have used

R has a built-in editor that may be sufficient for some users. More powerful editors are more productive for experienced programmers. All the editors below I have used my current favorite on Windows and Mac is RStudio.

- RStudio
- RWinEdt
- Tinn-R
- ESS (Emacs Speaks Statistics)
- Eclipse

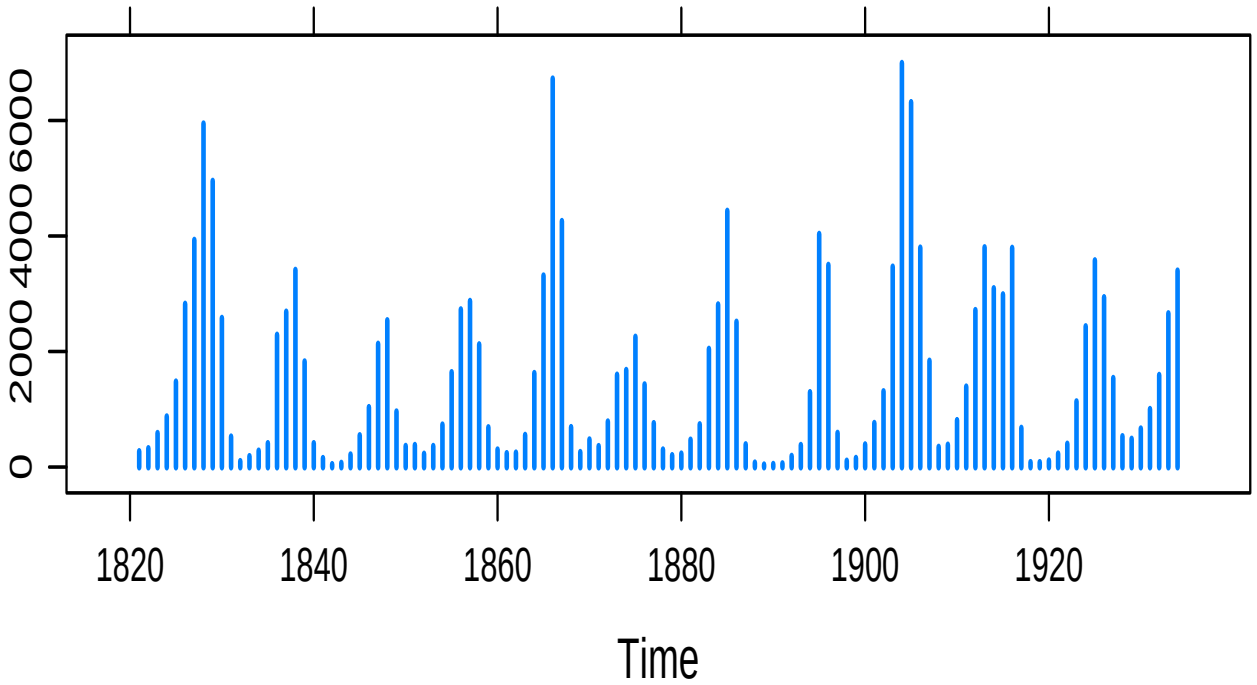
Time Series Plot

The famous lynx trapping time series.

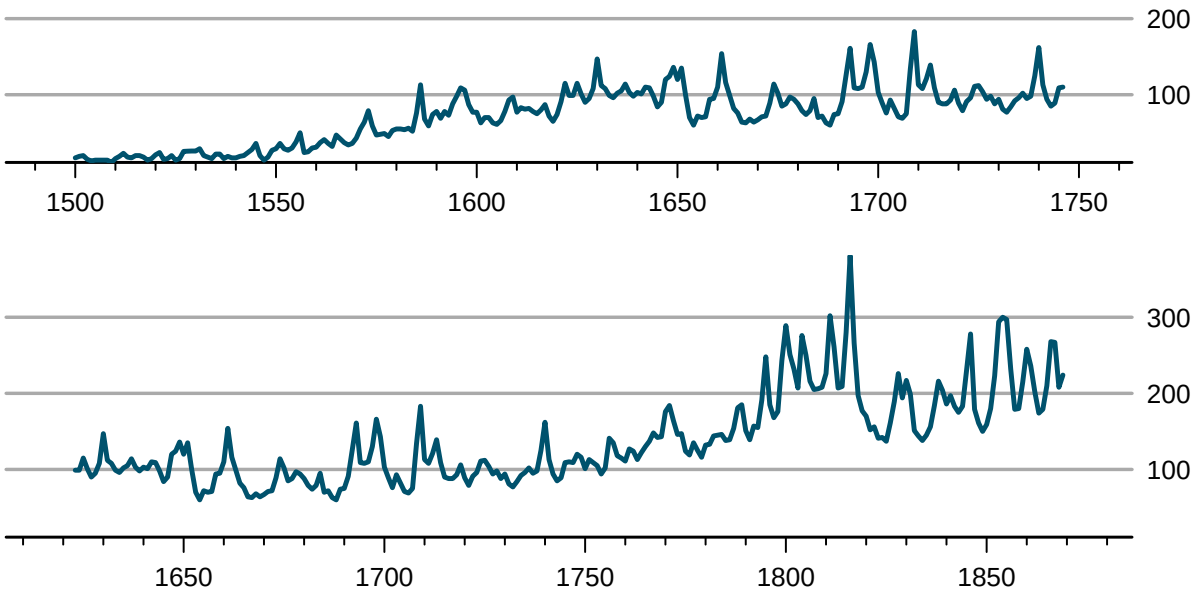


High density plot

The famous lynx trapping time series.



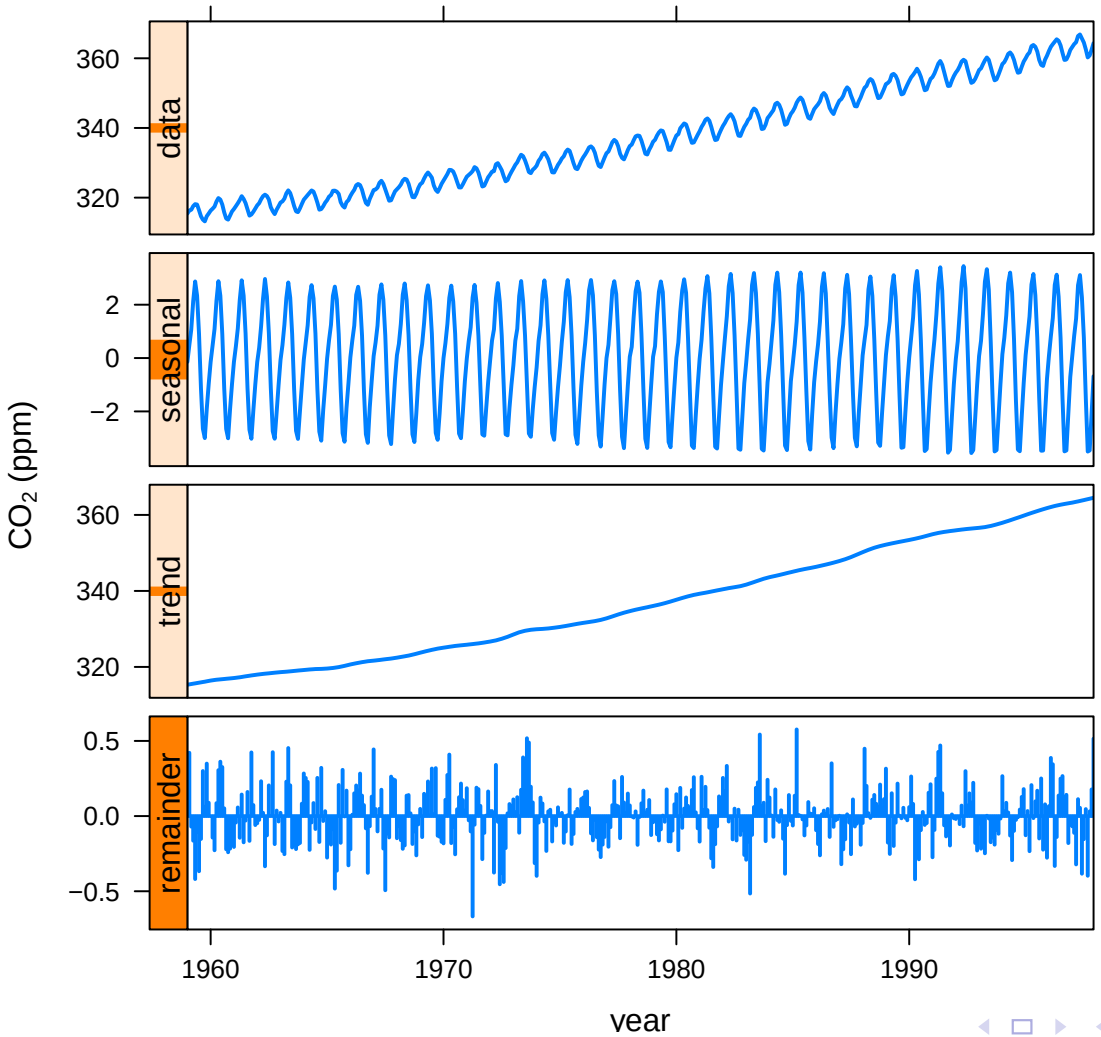
Cut-and-stack plot of Beveridge wheat price index



data source: tseries R package

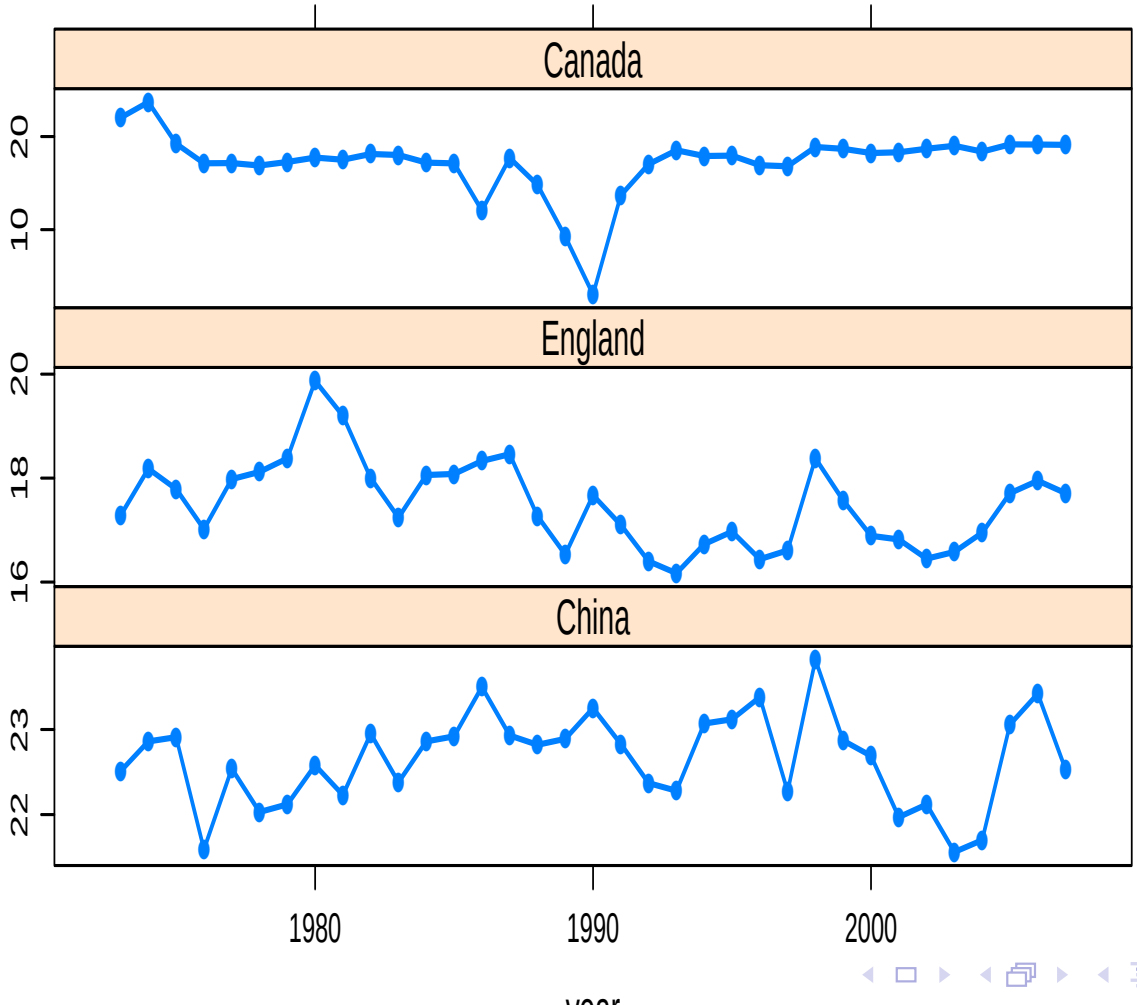
Lattice style - seasonal decomposition

Atmospheric CO_2 in Northern Hemisphere.

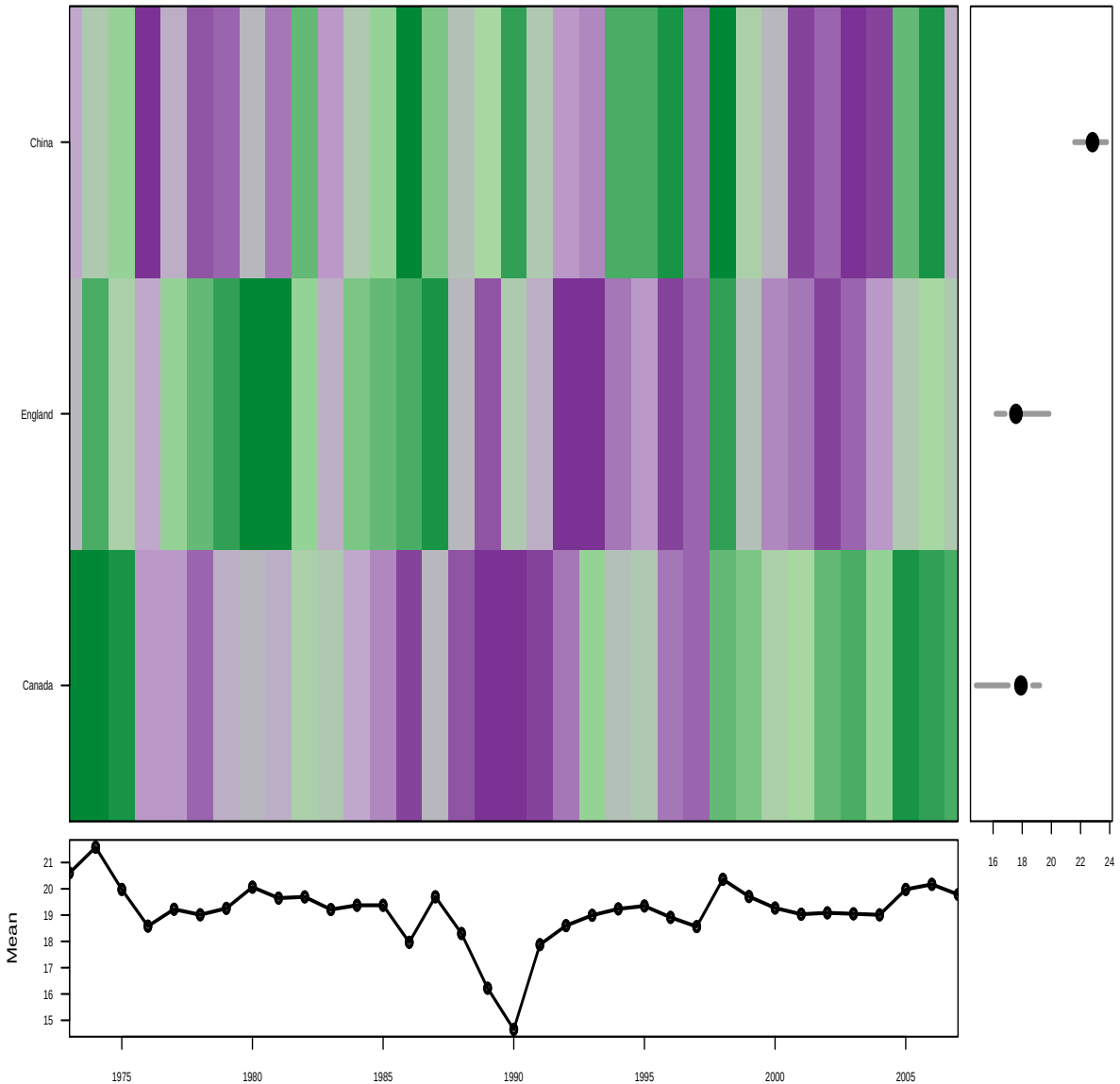


Lattice style - multiple time series

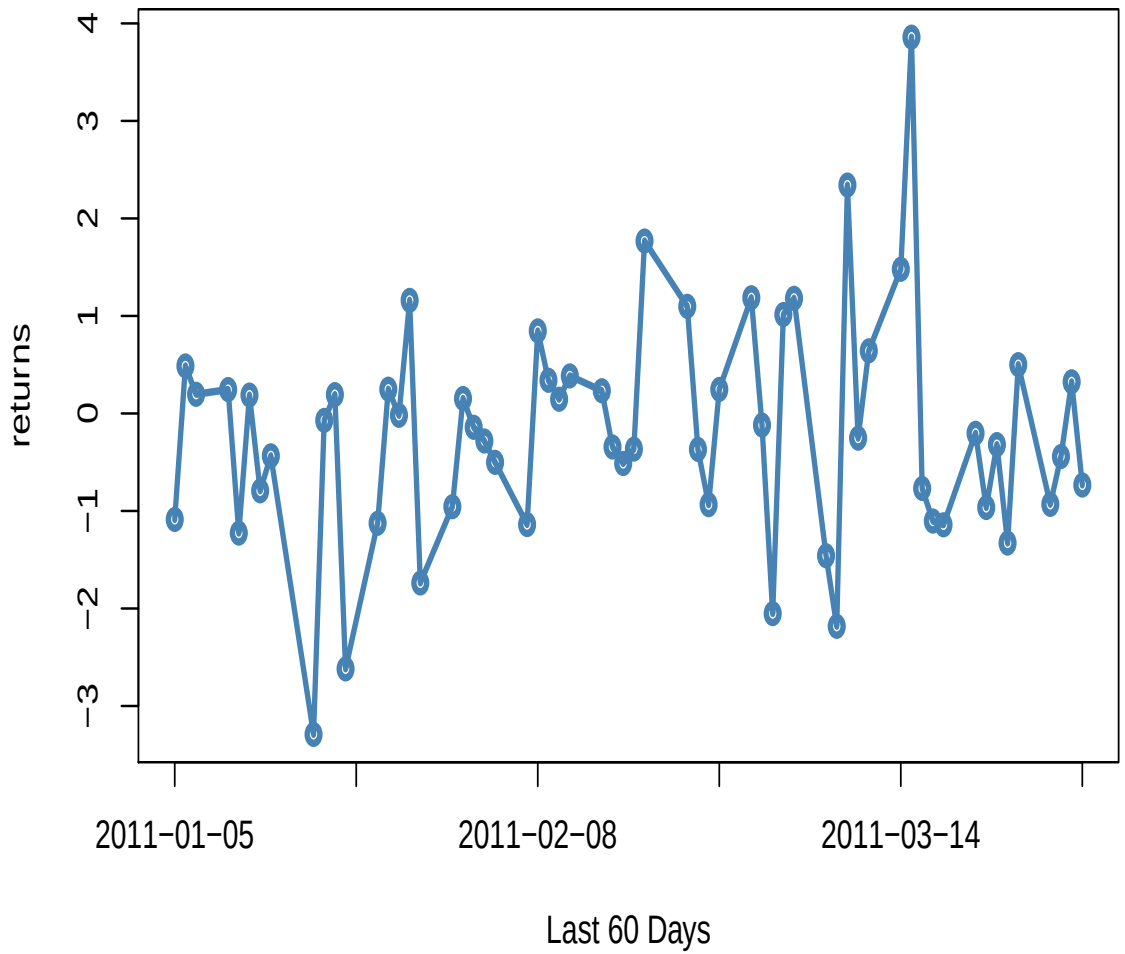
Mean annual temperature °C



High dimensional multiple time series



IBM, daily close, returns



R Base and Recommended Packages

Base and recommended packages are maintained by the R Core Team. Hence, these packages are of high quality. Other packages are maintained by individuals, so you have to more exercise caution. A complete list of time series functions and datasets available in these packages is given in the Appendices in our paper.

`http://cran.r-project.org/doc/FAQ/R-FAQ.html#R-Add_002dOn-Packages`

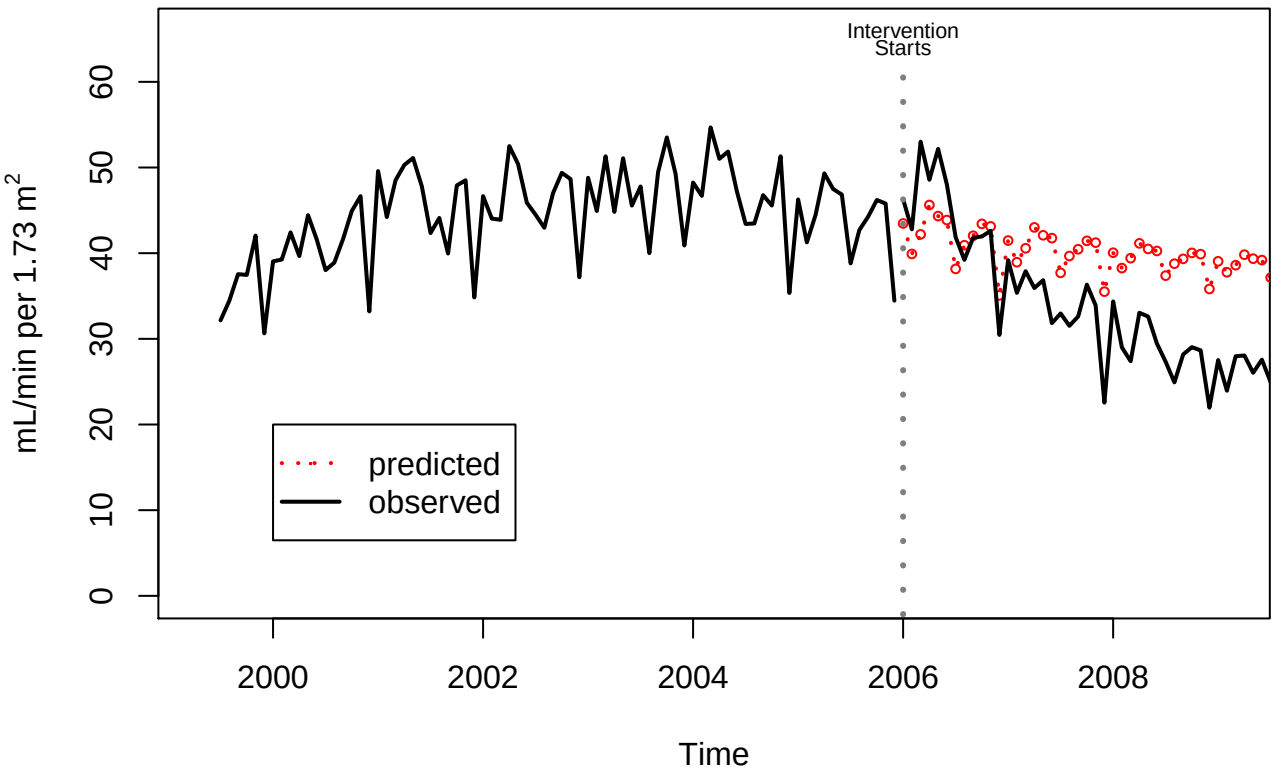
In our paper many other Packages that have been either published in *JSS or other referred or high-quality editorially reviewed books are discussed.*

ARIMA

- `arima()`
- `acf()`
- `tsdiag()`
- `Box.test()`
- `spectrum()`
- `ar()`
- `HoltWinters()`

`arima()` and `HoltWinters()` have methods functions `predict` and `plot`.

Intervention analysis with `arima()`



Structural Time Series Models

Structural time series models Harvey (1989) are also implemented using Kalman filtering in the function `StructTS()`. Since the Kalman filter is used, Kalman smoothing is also available and it is implemented in the function `tsSmooth()`. The basic structural model is comprised of an observational equation,

$$z_t = \mu_t + s_t + e_t, \quad e_t \sim NID(0, \sigma_e^2)$$

and the state equations,

$$\mu_{t+1} = \mu_t + \xi_t, \quad \xi_t \sim \mathrm{NID}(0, \sigma_\zeta^2),$$

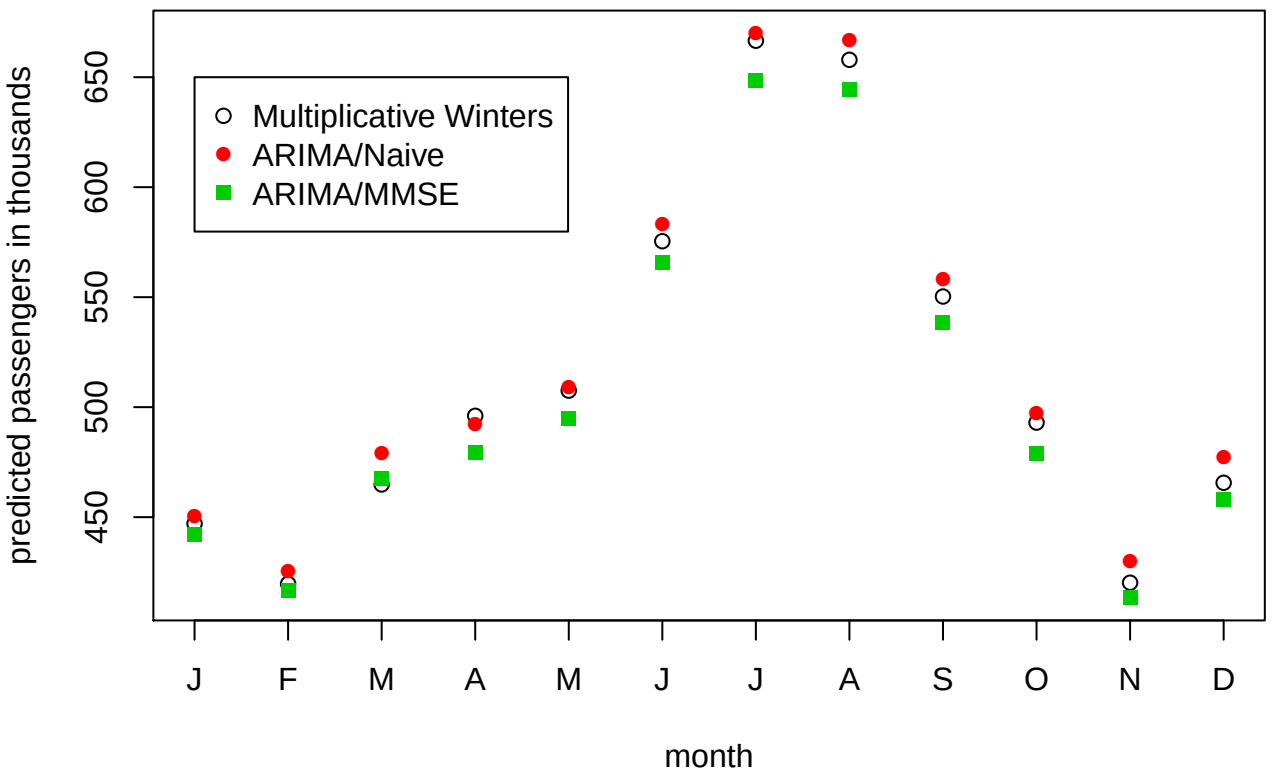
$$\nu_{t+1} = \nu_t + \zeta_t, \quad \zeta_t \sim \mathrm{NID}(0, \sigma_\zeta^2),$$

$$\gamma_{t+1} = -(\gamma_t + \ldots + \gamma_{t-s+2}) + \omega_t, \quad \omega_t \sim \mathrm{NID}(0, \sigma_\eta^2).$$

Special Structural Time Series Models

The local linear trend model is obtained by omitting the term involving γ_t in the observational equation and the last state equation may be dropped as well. Setting $\sigma_\zeta^2 = 0$ in the local linear trend model results in a model equivalent to the ARIMA(0,2,2). Setting $\sigma_\xi^2 = 0$ produces the local linear model which is also equivalent to the ARMA(0,1,1).

Forecast comparison: Airline passengers 1961



State space model

In general, the state space model is comprised of two equations, the observation equation:

$$\mathbf{y}_t = \mathbf{d}_t + \mathbf{Z}_t \boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t \tag{1}$$

and the state equation:

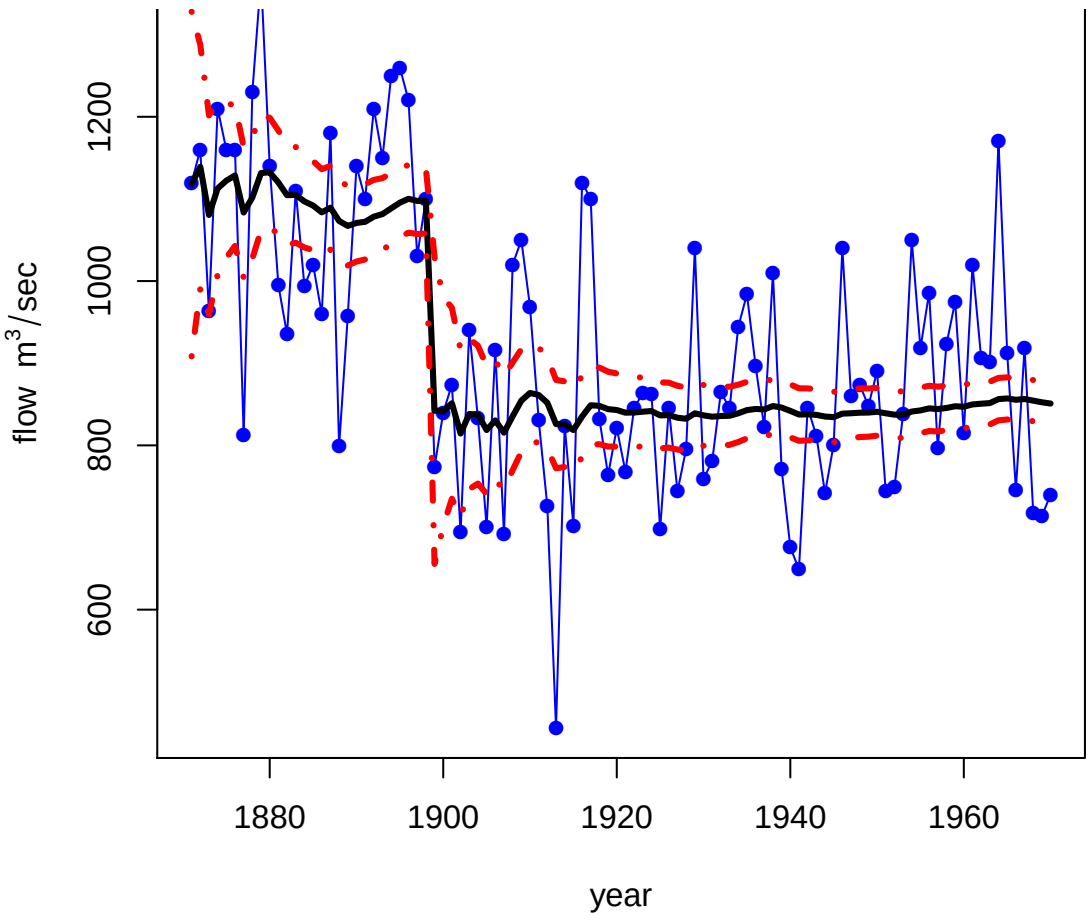
$$\boldsymbol{\alpha}_t = \mathbf{c}_t + \mathbf{T}_t \boldsymbol{\alpha}_{t-1} + \mathbf{R}_t \boldsymbol{\eta}_t, \tag{2}$$

where the white noises, $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$, are multivariate normal with mean vector zero and covariance matrices Q_t and H_t respectively. The white noise terms are uncorrelated, $\mathrm{E}\{\boldsymbol{\epsilon}_t' \boldsymbol{\eta}_t\} = 0$.

Tusell (2011) in a *JSS* paper reviews some of the R packages that implement the above model. Univariate and multivariate ARIMA models are special cases of this model.

Annual flows of Nile River

$$y_t = \theta_t + v_t, \qquad v_t \sim \mathcal{N}(0, V)$$
$$\theta_t = \theta_{t-1} + w_t, \quad w_t \sim \mathcal{N}(0, W)$$



Our R time series packages on CRAN

- **arfima**
- **portes**
- **ltsa**
- **FGN**
- **FitAR**
- **FitARMA**
- **mleur**
- **deseasonalize**
- **par**
- **Kendall**

Tsay Example

TAR and related models are also discussed by Tsay (2010) and some R scripts are provided as well the companion package **FinTS** that includes data sets from the book. Figure shows monthly U.S. unemployment. Tsay fits the two regime TAR model,

$$\begin{aligned} y_t &= 0.083y_{t-2} + 0.158y_{t-3} + 0.0118y_{t-4} - 0.180y_{t-12} + a_{1,t} \quad \text{if } y_{t-1} \leq \\ &= 0.421y_{t-2} + 0.239y_{t-3} - 0.127y_{t-12} + a_{2,t} \quad \text{if } y_{t-1} > 0.01, \end{aligned}$$

where y_t is the differenced unemployment series. The estimated standard deviations of $a_{2,t}$ and $a_{2,t}$ were 0.180 and 0.217. Tsay remarks that the TAR provides more insight into the time-varying dynamics of the unemployment rate than the ARIMA.

GARCH Models

Volatility refers to the random and autocorrelated changes in variance exhibited by many financial time series. The GARCH family of models capture quite well volatility clustering as well as the thick-tailed distributions often found with financial time series such as stock returns and foreign exchange rates.
A GARCH(p, q) sequence $a_t, t = \ldots, -1, 0, 1, \ldots$ is of the form

$$a_t = \sigma_t \epsilon_t$$

and

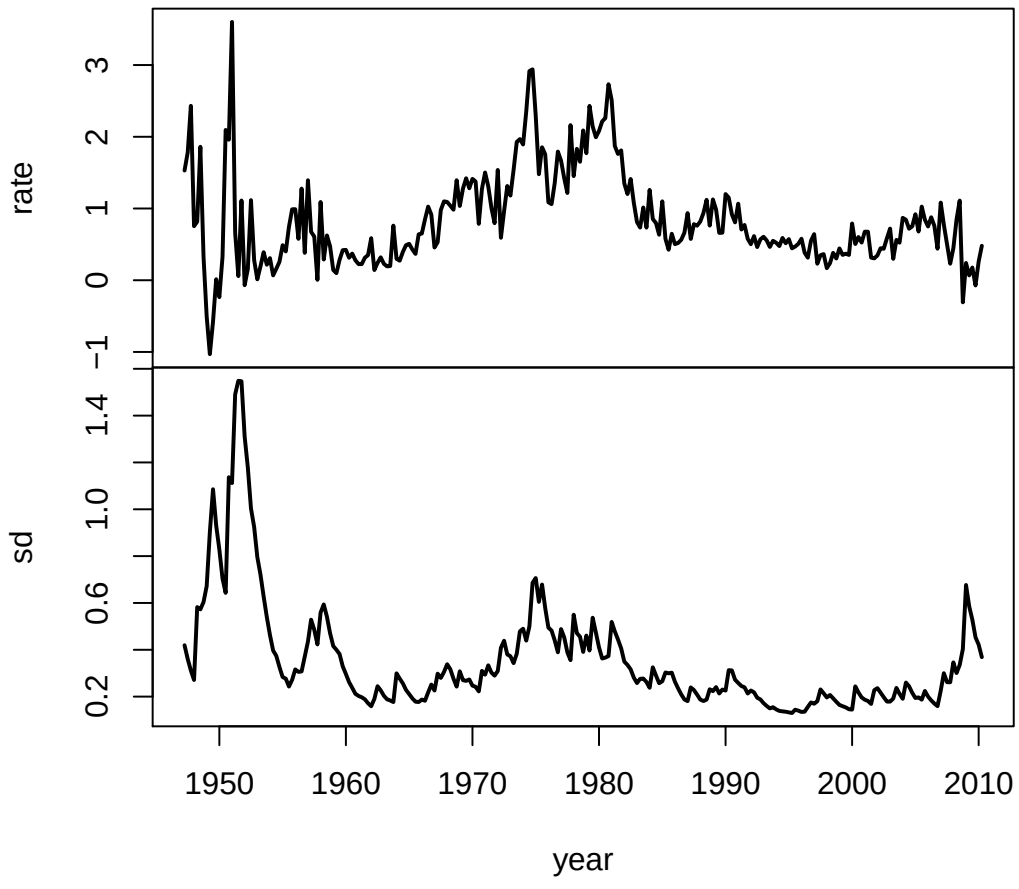
$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_j a_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2,$$

where $\alpha_0 > 0,\ \alpha_i \geq 0,\ 1 \leq i \leq p,\ \beta_j \geq 0,\ 1 \leq j \leq q$ are parameters. The errors ϵ_t are assumed to be independent and identically distributed from a parametric distribution such as normal, generalized error distribution (GED), Student-t or skewed variations of these distributions.

U.S. inflation

We used the GNP deflator for 1947-01-01 to 2010-04-01. There were $n = 254$ observations which are denoted by $z_t, t = 1, \ldots, n$. Then the inflation rate may be estimated by the logarithmic difference, $r_t = \log(z_t) - \log(z_{t-1})$. The following ARMA/GARCH model was fit using the function `garchFit()` in **fGarch**, $r_t = 0.103 + 0.369 r_{t-1} + 0.223 r_{t-2} + 0.248 r_{t-3} + \epsilon_t$, and $\sigma_t^2 = 0.004 + 0.269 \epsilon_{t-1}^2 + 0.716 \sigma_{t-1}^2$.

U.S. unemployment rate, seasonally adjusted, January 1948 to March 2004.

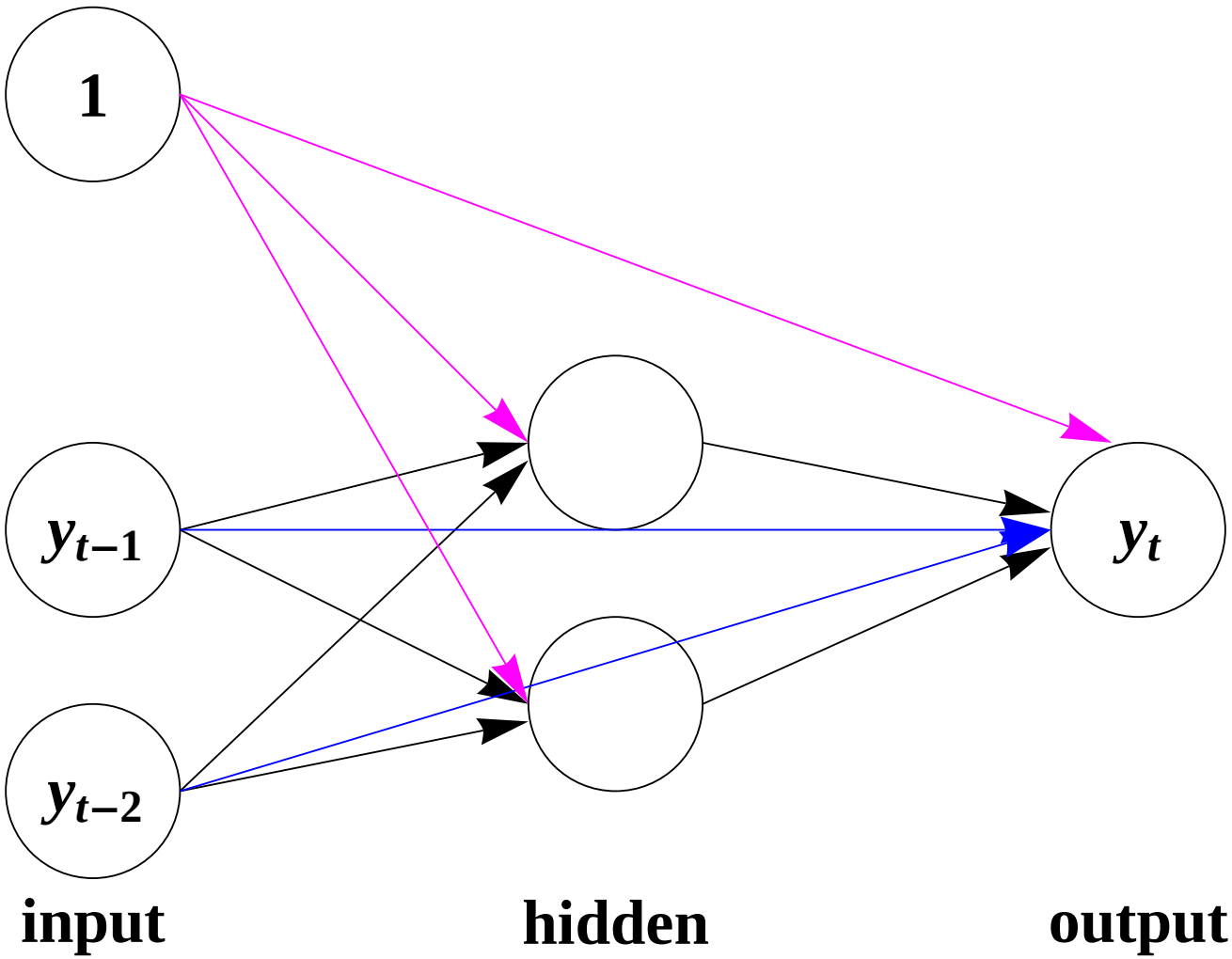


Feed-forward neural nets - nonlinear AR(2)

$$y_t = f_o\left(a + \sum_{i=1}^{p} \Omega_i x_i + \sum_{j=1}^{H} w_j f\left(\alpha_j + \sum_{i=1}^{p} \omega_{i,j} x_{t-i}\right)\right), \qquad (3)$$

where $\hat{y}_t$ is the predicted time series at time t and $y_{t-1}, \ldots, y_{t-p}$ are the lagged inputs, f_o is the activation function for the output node, f is the activation function for each of the H hidden nodes, $\omega_{i,j}$ are the p weights along the connection for the j-th hidden node, Ω_i is the weight in the skip-layer connection, and a is the bias connection. There are $m(1 + H(p + 2))$ unknown parameters that must be estimated. The hyperparameter H, is determined by cross-validation. The activation functions f and f_o are often logistic.

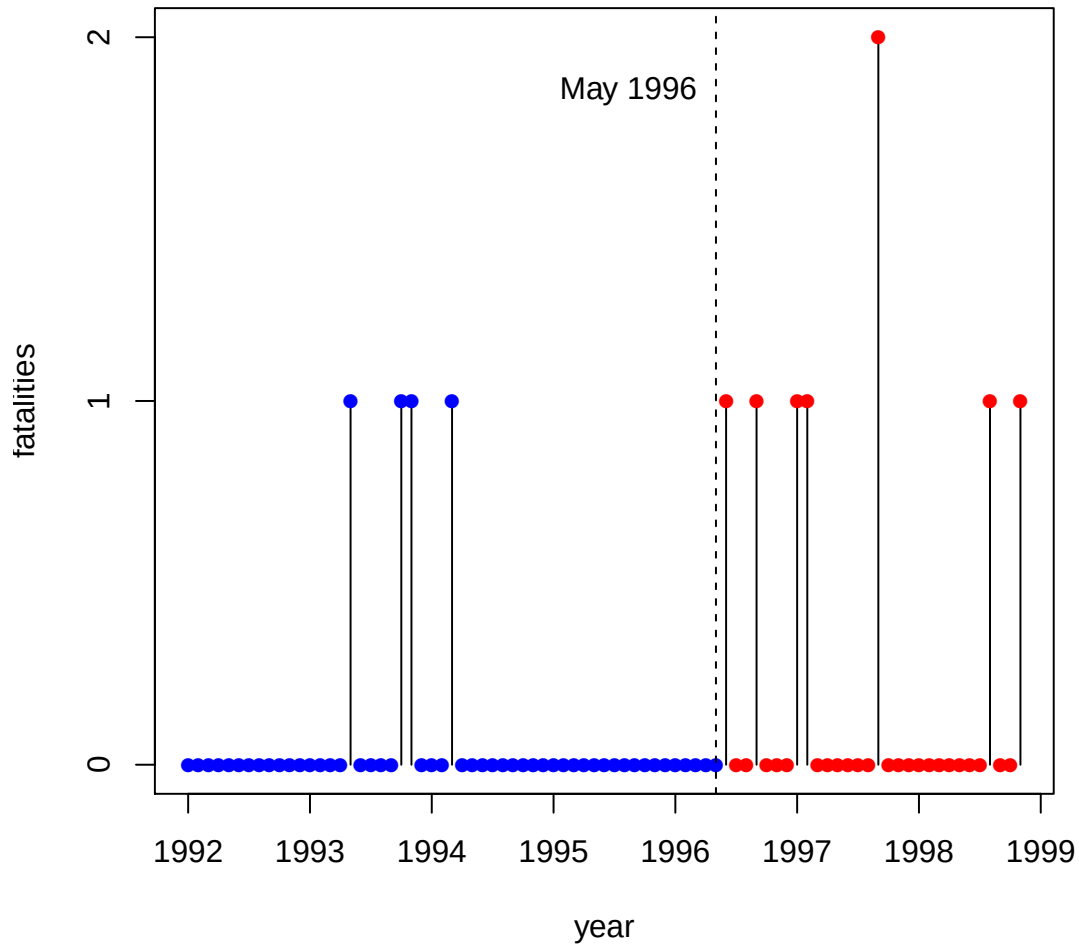
Nonlinear AR(2)



Remarks

- easily generalized for multivariate time series
- equivalent to project pursuit regression
- **nnet**
- **RWeka**

Late night automobile accident fatalities



Fitted GLM model

The total fatalities per month, y_t, are Poisson distributed with mean μ_t, where $\hat{\mu}_t = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_t + \hat{\beta}_2 y_{t-1}\}$, $\hat{\beta}_0 \doteq -2.54$, $\hat{\beta}_1 \doteq 1.17$, and $\hat{\beta}_2 \doteq -0.07$. There is no evidence of lagged dependence but the intervention effect, β_2 is significant with $p < 0.10$.

Further nonlinear models for time series

- generalized additive models
- MARS
- spline regression for trend estimation
- nearest neighbour prediction
- local linear
- `optim` and **maxLik**

DWT

Consider a time series of dyadic length, $z_t, t = 1, \ldots, n$, where $n = 2^J$. The discrete wavelet transformation (DWT) decomposes the time series into J wavelet coefficients vectors, $W_j, j = 0, \ldots, J-1$ each of length $n_j = 2^{J-j}, j = 1, \ldots, J$ plus a scaling coefficient V_J. Each wavelet coefficient is constructed as a difference of two weighted averages each of length $\lambda_j = 2^{j-1}$. Like the discrete Fourier transformation, the DWT provides an orthonormal decomposition, $W = \mathcal{W}Z$, where $W' = (W_1', \ldots, W_{J-1}', V_{J-1}')$, $Z = (z_1, \ldots, z_n)'$ and $\mathcal{W}$ is an orthonormal matrix.

DWT Algorithm

In practice, the DWT is not computed using matrix multiplication but much more efficiently using filtering and downsampling (Percival and Walden, 2000). The resulting algorithm is known as the pyramid algorithm and computationally it is even more efficient than the fast Fourier transform. Applying the operations in reverse order yields the inverse DWT. Sometimes a partial transformation is done.

Wavelet coefficients

The wavelet coefficients are associated with changes in the time series over the scale $\lambda_j = 2^{j-1}$ while the scaling coefficients, V_{J_0}, are associated with the average level on scale $\tau = 2^{J_0}$.

MODWT

The maximum overlap DWT or MODWT omits the downsampling. The MODWT has many advantages over the DWT even though it does not provide an orthogonal decomposition.

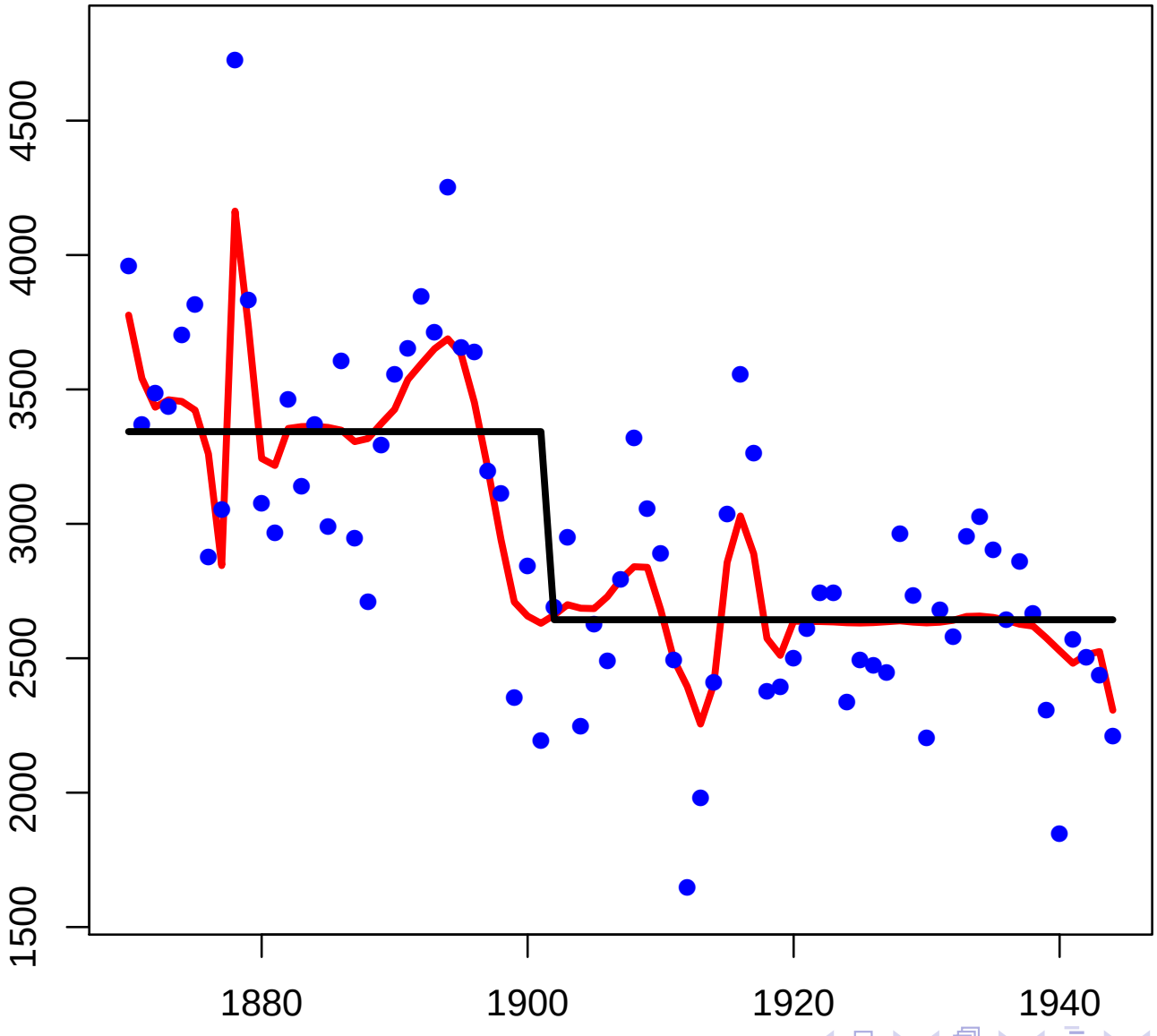
Books

Percival and Walden (2000) provide an extensive treatment of wavelet methods for time series research with many interesting scientific time series. Gencay (2002) follows a similar approach to wavelets but with an emphasis on financial and economic applications. There are several R packages that implement the methods discussed in these books.

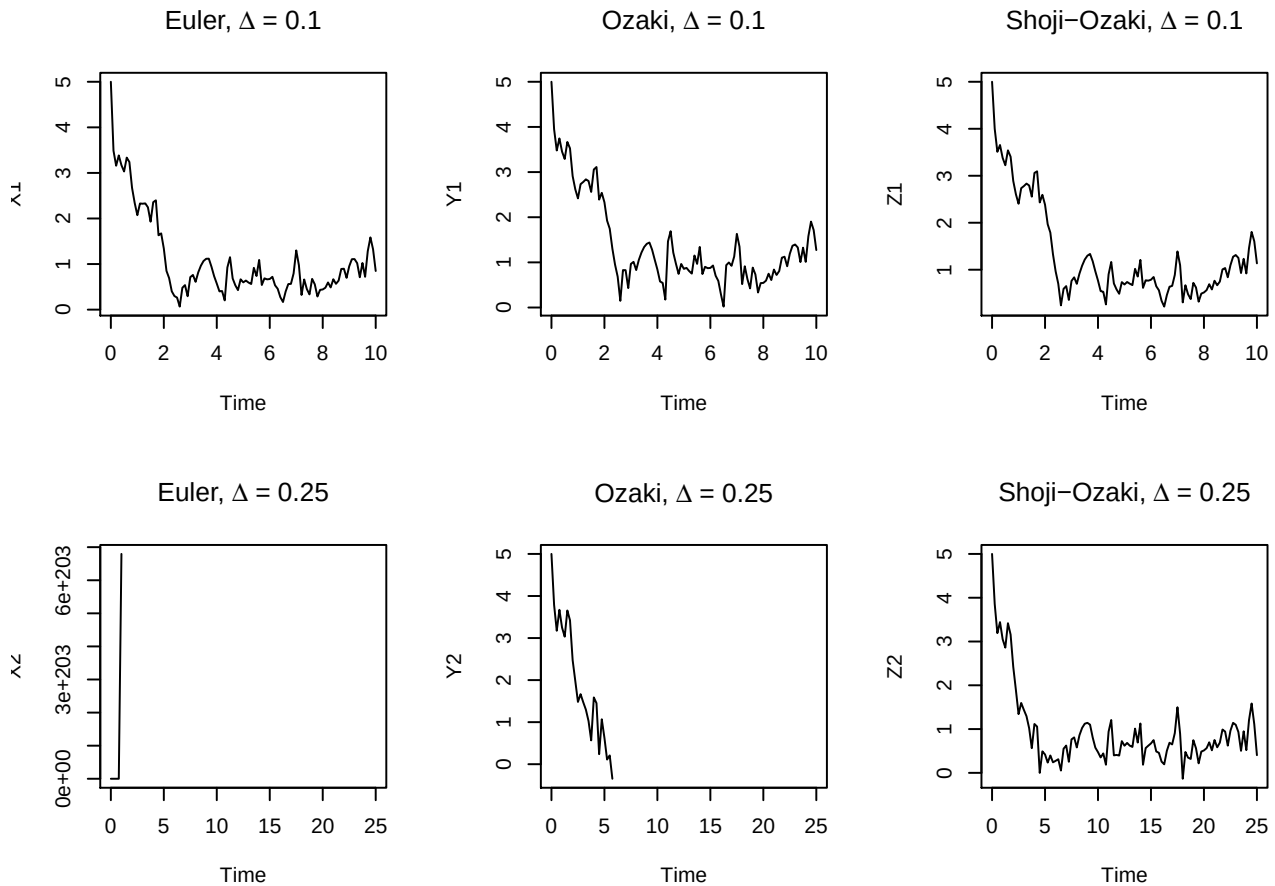
Annual Nile River Flows

We computed the denoised annual Nile riverflows using the universal threshold with hard thresholding and Haar wavelets and compared this to the fit a step intervention analysis time series model with AR(1) noise. The fitted step intervention is represented by the three line segments while the denoised flows are represented by the jagged curve.

Mean annual Nile flow, October to September, Aswan



Simulations of $dX(t) = (5 - 11x + 6x^2 - x^3)dt + dW(t)$



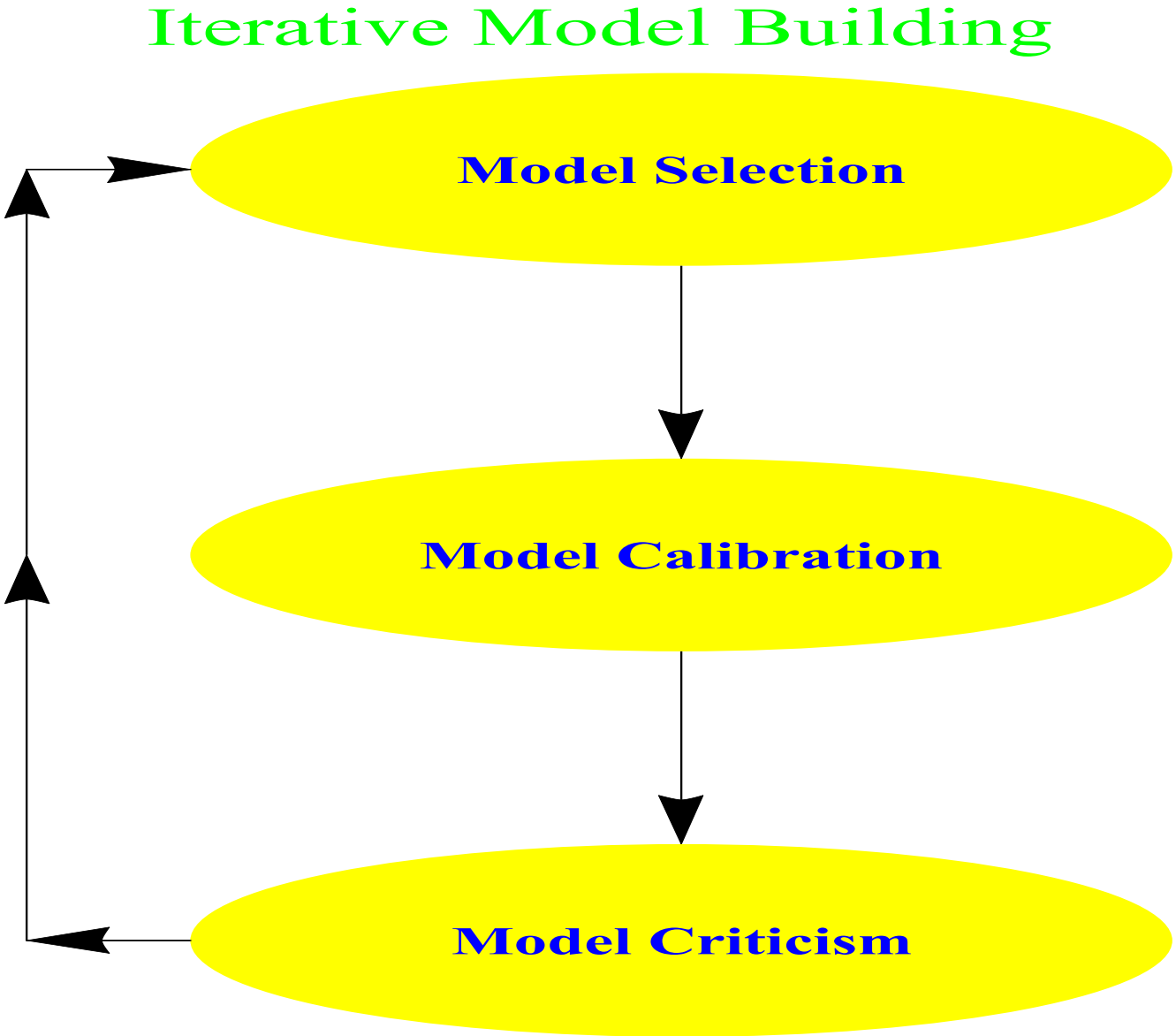
Three unit root models

$$\Delta Z_t = \beta_0 + \beta_1 t + \gamma Z_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta Z_{t-i} + e_t \quad (4)$$

$$\Delta Z_t = \beta_0 + \gamma Z_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta Z_{t-i} + e_t, \quad (5)$$

$$\Delta Z_t = \gamma Z_{t-1} + \sum_{i=1}^{p-1} \delta_i \Delta Z_{t-i} + e_t, \quad (6)$$

Iterative model building



Test statistic and its distribution

For all three models, the unit-root test is equivalent to testing $\mathcal{H}_0 : \gamma = 0$ is

$$\tau_i = \frac{\hat{\phi} - 1}{\mathrm{SE}(\hat{\phi})}, \quad i = 1, 2, 3,$$

where i denotes the model (6), (5), or (4) respectively. The distribution of τ_i has been obtained by Monte-Carlo simulation using response surface regression methods.

Test for stationarity

The KPSS test (Kwiatkowski et al. 1992) was developed in econometrics but is often used as an all round test for trend in environmental time series.
This test detects non-stationarity in the annual Nile riverflow series.

Theory for KPSS Test

The underlying model for the KPSS level test:

$$\begin{aligned} y_t &= \mu_t + N_t \\ \mu_t &= \mu_{t-1} + \xi_t, \end{aligned}$$

where N_t is assumed to be a stationary and possibly heteroscedastic time series and $xi_t \sim \mathrm{NID}(0, \sigma_\xi^2)$.
The test statistic is derived as a Lagrange multiplier test for $\mathcal{H}_0 : \sigma_\xi^2 = 0$.

Cointegration

In the simplest case, two time series that are both difference-stationary are said to be cointegrated when a linear combination of them is stationary. Some classic examples:

- consumption and income
- wages and prices
- short and long term interest rates